A Dynamic Subspace Method for Hyperspectral Image Classification

Jinn-Min Yang, Bor-Chen Kuo, Pao-Ta Yu, *Member, IEEE*, and Chun-Hsiang Chuang

***Abstract*—Many studies have demonstrated that multiple classifier systems, such as the random subspace method (RSM), obtain more outstanding and robust results than a single classifier on extensive pattern recognition issues. In this paper, we propose a novel subspace selection mechanism, named the dynamic subspace method (DSM), to improve RSM on automatically determining dimensionality and selecting component dimensions for diverse subspaces. Two importance distributions are proposed to impose on the process of constructing ensemble classifiers. One is the distribution of subspace dimensionality, and the other is the distribution of band weights. Based on the two distributions, DSM becomes an automatic, dynamic, and adaptive ensemble. The real data experimental results show that the proposed DSM obtains sound performances than RSM, and that the classification maps remarkably produce fewer speckles.**

***Index Terms*—Kernel smoothing (KS), random subspace method (RSM), small sample size (SSS) classification.**

I. Introduction

IN hyperspectral imaging, data from the new generation sensors consist of a large number of spectral bands that provide the potential to improve the discrimination of objects. However, one of the difficulties for supervised classification inhibiting this potential is the constraint of training sample size because the ground truth is generally expensive and difficult to acquire. Therefore, we have to face the small sample size (SSS) problem, that is, the number of available training samples is much smaller than the dimensionality. Under this circumstance, the generalization ability of the resulting classifier is weak, and the variances of its classification results are large [1], [2]. In other words, the classifier suffers from the well-known Hughes phenomenon [3] or the curse of dimensionality [4] in classification results.

The random subspace method (RSM) proposed by Ho [5], [6] is one of the multiple classifier systems, providing a way of alleviating sample size and high-dimensionality concerns. It is a general technique that can be used with any type of base classifier [7]–[12]. Moreover, much research [13]–[15] has demonstrated its validness for hyperspectral image classification. In RSM, each weak classifier is constructed in a subspace with bands randomly selected from the original ones, and the subspace dimensionality is usually predefined. Then, a final decision rule of weak classifiers is obtained by a simple majority vote. However, there are two inadequacies in RSM. One is that the dimensionality of subspace is not clearly defined, and the other is its random rule for selecting bands.

Ho suggests that desirable results are obtained by setting the dimensionality of subspace to approximately half of the dimensionality of original space [5], [16]. This result is based on the decision tree classifier, but it may not be extended to all kinds of classifiers. For instance, the suitable dimensionality of subspaces for a maximum likelihood (ML) classifier depends on the size of the training samples. The question of how to choose a suitable subspace size for the employed classifier will then arise. In addition, the random strategy assumes that the selected probability of each band to form a subspace is the same, but the discriminating power of each band is actually different.

In this paper, we propose the dynamic subspace method (DSM) for constructing component classifiers with adaptive subspaces to adjust the shortcomings of RSM. DSM works on the basis of two major distributions, namely, W and R, denoting the distributions of band weights and subspace dimensionality, respectively. The component bands to form the subspace are selected with the probability based on the W distribution, and the number of selected bands is automatically determined based on the R distribution. In fact, the R distribution records the importance of all possible subspace size, which is estimated by the kernel density estimation technique [17], [18] on some resubstitution performances of partial dimensionalities. Most importantly, it would be updated in the training process of constructing DSM. Comparing to a heuristic search or manual method, this scheme shows its dynamic selection manner for deciding the applicable dimensionality with respect to the employed classifiers.

Recently, the classification technique integrating both spectral and spatial information has rapidly developed for the hyperspectral image classification [19]–[22], where the Markov random field (MRF) is one of the popular models to exploit the spatial context between neighboring pixels in an image. In this study, an MRF-based contextual classification [23] is also applied to the proposed DSM as the base learner because it suffers from the SSS problems.

Manuscript received April 2, 2009; revised August 12, 2009, December 7, 2009, and January 12, 2010. Date of publication April 5, 2010; date of current version June 23, 2010. This work was supported in part by the National Science Council, Taiwan, under Grant NSC 98-2221-E-142-005.

J.-M. Yang is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan, and also with the Department of Mathematics Education, National Taichung University, Taichung 403, Taiwan.

B.-C. Kuo is with the Graduate School of Educational Measurement and Statistics, National Taichung University, Taichung 40306, Taiwan (e-mail: kbc@mail.ntcu.edu.tw).

P.-T. Yu is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan.

C.-H. Chuang is with the Institute of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2010.2043533

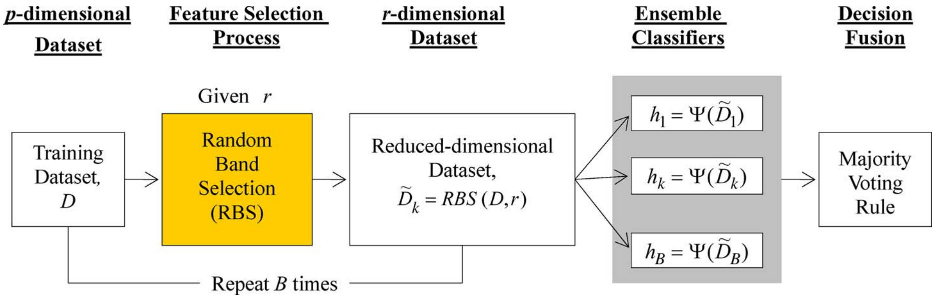

p-dimensional Dataset
Feature Selection Process
r-dimensional Dataset
Ensemble Classifiers
Decision Fusion
Given r
Training Dataset, D
Random Band Selection (RBS)
Reduced-dimensional Dataset, $\widetilde{D}_k = RBS(D, r)$
$h_1 = \Psi(\widetilde{D}_1)$
$h_k = \Psi(\widetilde{D}_k)$
$h_B = \Psi(\widetilde{D}_B)$
Majority Voting Rule
Repeat B times

Fig. 1. Framework of producing ensemble classifiers by RSM.

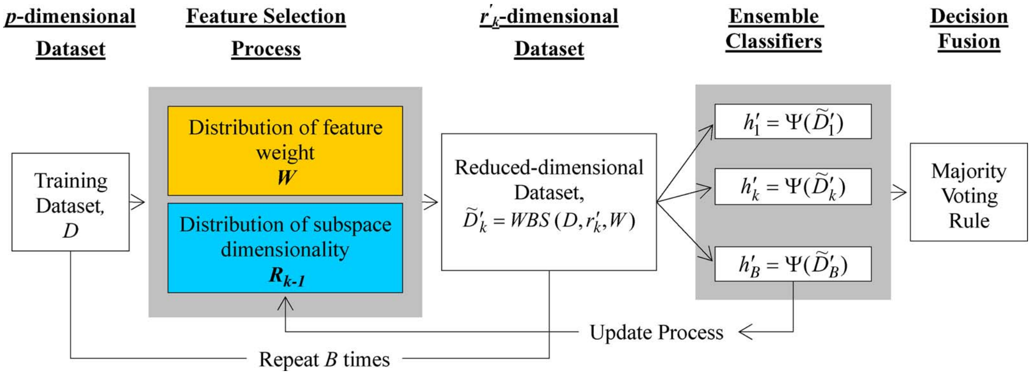

p-dimensional Dataset
Feature Selection Process
r'_k-dimensional Dataset
Ensemble Classifiers
Decision Fusion
Training Dataset, D
Distribution of feature weight W
Distribution of subspace dimensionality R_{k-1}
Reduced-dimensional Dataset, $\widetilde{D}'_k = WBS(D, r'_k, W)$
$h'_1 = \Psi(\widetilde{D}'_1)$
$h'_k = \Psi(\widetilde{D}'_k)$
$h'_B = \Psi(\widetilde{D}'_B)$
Majority Voting Rule
Update Process
Repeat B times

Fig. 2. Famework of producing ensemble classifiers by DSM, where the update process is to estimate the resubstitution accuracy of h'_k as the feedback to update the R distribution.

The rest of this paper is organized as follows. A brief review of the RSM will be described in Section II. New methods will be derived in Section III. For evaluating the performance of the proposed method, real hyperspectral image data experiments are designed in Section IV, and the experimental results are reported in Section V. Section VI contains some comments and conclusions.

II. RSM

The RSM proposed by Ho [5], [6] is an ensemble technique based on random band selection (RBS). Let $D = \{(\mathbf{x}_i, c_i) | 1 \leq i \leq N\}$ represent the original p-dimensional data set that is composed of N training samples, where $\mathbf{x}_i \in \Re^p$ with class label $c_i \in C = \{1, 2, \ldots, L\}$, and L is the total number of classes. In RSM, given a predefined subspace dimensionality $r < p$, the RBS process randomly selects r bands from the original p-dimensional space such that D reduces to $\widetilde{D} = RBS(D, r) = \{(\widetilde{\mathbf{x}}_i, c_i) | 1 \leq i \leq N\}$, where $\widetilde{\mathbf{x}}_i \in \Re^r$; then, $\widetilde{D}$ returns as an input to the learning algorithm Ψ, which outputs a classifier $h = \Psi(\widetilde{D})$. This process will repeat B times to construct ensemble classifiers $H = \{h_1, h_2, \ldots, h_B\}$. In the classification procedure, diverse class labels of a test sample Y are obtained by these classifiers and then combined together by simple majority voting to obtain a final decision $F = \arg\max_{c \in \{1,2,\ldots,L\}} card(k | h_k(Y) = c, k = 1, 2, \ldots, B)$, where $card(A)$ denotes the cardinality of the set A. The framework of RSM is shown in Fig. 1, where $\widetilde{D}_k$ denotes the kth reduced-dimensional data set of the original data set D.

The RSM has been theoretically and experimentally proven to be beneficial for the SSS problem, but there are still two prominent weaknesses that need to be improved. One is that the dimensionality of subspace is fixed and needs to be predefined, generally selected by the trial-and-error method; the other is that its randomized band selection mechanism makes the equally selected probabilities of informative and noninformative bands. In Section III, a novel method of the multiple classifier system, named DSM, will be proposed to overcome these weaknesses of RSM.

III. DSM

In this section, DSM is introduced, and how the drawbacks of RSM are overcome is shown. The design of DSM is displayed in Fig. 2, where two innovative distributions, namely, W and R, are imposed in the process of subspace selection. In addition, $\widetilde{D}'_k$ and r'_k represent the kth reduced-dimensional data set of D and its corresponding dimensionality, respectively. Compared to RSM, the contributions of bands are assumed differently, that is, band selection is no longer according to the uniform distribution. We propose the importance distribution of band weight W to model the probability of bands being selected. Importantly, the subspace dimensionality is neither predefined nor a fixed number but is drawn from the importance distribution of subspace dimensionality R. An update process for R is also proposed in each overproduction. DSM constructs $\widetilde{D}'_k$ with r'_k bands based on W and R distributions. The algorithm of DSM is summarized in Algorithm 1.

In the following, W and R distributions are defined, respectively, and the DSM algorithm is explained.

A. W Distribution

The design of W distribution is based on the principle that beneficial bands carry larger probabilities to be selected, and

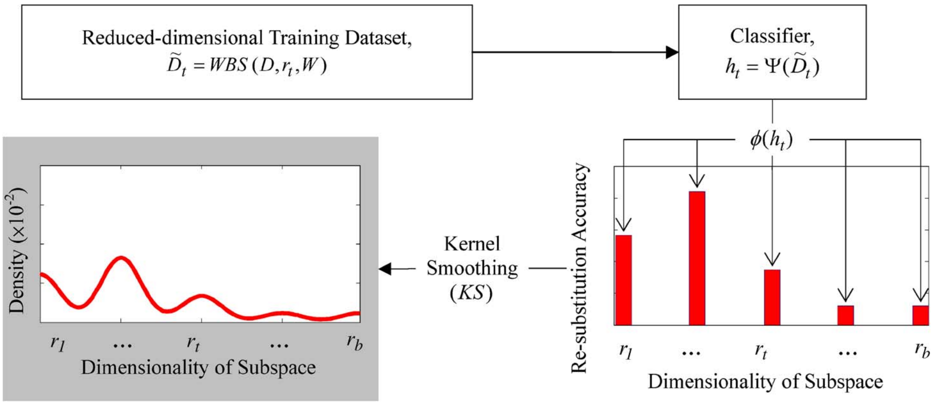

Reduced-dimensional Training Dataset,
$\widetilde{D}_t = WBS(D, r_t, W)$
Classifier,
$h_t = \Psi(\widetilde{D}_t)$
$\phi(h_t)$
Re-substitution Accuracy
r_1 … r_t … r_b
Dimensionality of Subspace
Kernel Smoothing (KS)
Density ($\times 10^{-2}$)
r_1 … r_t … r_b
Dimensionality of Subspace

Fig. 3. Initial procedure for estimating the R_0 distribution.

smaller probabilities are given to the futile ones. A class-based band selection for creating an ensemble of classifiers was proposed in [24]. It is time consuming and not suitable for our DSM. Hence, two simple and multiclass-based band selection methods are proposed for DSM. The band selection processes are based on two W distributions, $W_{ACC(\Psi)}$ and W_{LDA}, where the subindices $ACC(\Psi)$ and LDA represent the resubstitution accuracy by applying the classifier Ψ and the class separability of Fisher's linear discriminate analysis (LDA) [25], respectively. Note that the histogram approach [18] is utilized for density estimation of both distributions. The following are their formulations, and the procedure of selecting bands based on the W distribution is also introduced.

1) $W_{ACC(\Psi)}$ Distribution: The $W_{ACC(\Psi)}$ distribution is built according to the so-called resubstitution accuracy [25], which is the classification accuracy of training data. In this paper, the resubstitution accuracy is obtained by applying the base classifier Ψ to each individual band. Assume that $W_{ACC(\Psi)}$ is a random variable with a probability mass function (pmf) given by a probability vector $f_{W_{ACC(\Psi)}} = (f_{W_{ACC(\Psi)}}(1), f_{W_{ACC(\Psi)}}(2), \ldots, f_{W_{ACC(\Psi)}}(p))$, where

$$f_{W_{ACC(\Psi)}}(j) = \frac{\phi_j}{\sum\limits_{k=1}^{p} \phi_k}, \qquad j = 1, 2, \ldots, p. \tag{1}$$

ϕ_j denotes the resubstitution classification accuracy by applying the base classifier Ψ to the jth band only.

2) W_{LDA} Distribution: Another measurement used to assign weight to individual bands in this study is based on the class separability of Fisher's LDA [25], which is referred to the power of discrimination and is measured by $J = \mathrm{tr}(S_w^{-1} S_b)$. The value of J, computed by the trace of the inverse of the within-class scatter matrix (S_w) times the between-class scatter matrix (S_b), should be large to the beneficial bands but small to the futile ones. Assume that W_{LDA} is a random variable with pmf given by a probability vector $f_{W_{\mathrm{LDA}}} = (f_{W_{\mathrm{LDA}}}(1), f_{W_{\mathrm{LDA}}}(2), \ldots, f_{W_{\mathrm{LDA}}}(p))$, where

$$f_{W_{\mathrm{LDA}}}(j) = \frac{J_j}{\sum\limits_{k=1}^{p} J_k}, \quad J_j = \mathrm{tr}\left(S_{wj}^{-1} S_{bj}\right), \quad j = 1, 2, \ldots, p. \tag{2}$$

Note that J_j denotes the discrimination power of the jth spectral band.

3) Band Selection Based on the W Distribution: The foundation of the band selection algorithm based on the W distribution (W_{ACC}, W_{LDA}, or the uniform distribution) is the theory of pseudorandom number generation [26]. The inversion method of the pseudorandom number generation is used to implement the algorithm for selecting the desired bands. Assume that there are r bands that need to be selected; the steps of selecting these bands are described as follows.

1) Generate a uniform random number υ on [0, 1].
2) Select the kth band if $F_W(k-1) < \upsilon < F_W(k)$, where F_W denotes the cumulate density function of the W distribution, and $1 \le k \le p$.
3) Set the $f_W(k) = 0$ and renormalize the W distribution.
4) Go back to Step 1 until r bands have been selected.

Finally, a reduced-dimensional data set $\widetilde{D} = WBS(D, r, W)$ is obtained, where "WBS" denotes the acronym of W-based band selection.

B. R Distribution

The function of R distribution is to indicate how many dimensions are suitable for the employed base classifier. The procedure to establish R distribution includes two steps. First, we build R_0, an initial distribution of R, by applying Ψ to b different dimensional data sets with dimensionalities $r_1, \ldots, r_b$. Second, the kernel smoothing (KS) density estimation [17], [18] (or "Parzen density estimation" [25]) is utilized to smoothen R_0, making it a continuous one. Fig. 3 illustrates the aforementioned procedure. KS is an important and popular nonparametric technique for which prior knowledge about the functional form of the conditional probability distributions is not available or is not used explicitly [27].

As shown from the left plot in Fig. 3, the b training data set for building R_0 is generated based on W, i.e., $\widetilde{D}_t = WBS(D, r_t, W)$, $t = 1, 2, \ldots, b$, where the dimensionality r_t is given by

$$r_t = 1 + \left[(t-1) \times \frac{(p-1)}{(b-1)} \right]. \tag{3}$$

Next, we need to compute the resubstitution classification accuracy $\phi(h_t)$ by applying Ψ to $\widetilde{D}_t$. Then, the R_0 distribution

TABLE I
DESCRIPTION OF ALGORITHMS USED FOR COMPARISON

Algorithm	**Description**
Single Classifier	Using only a single classifier without any dimension reduction
RSM	Original RSM using half of the original space size for subspace
DSM	Dynamic subspace method with random band selection
DSMw1	DSM with the re-substitution accuracy as the band weights
DSMw2	DSM with the separability of Fisher's LDA as the band weights

is built. Finally, we get the continuous R_0 distribution by KS by

$$f_R(r) = \frac{1}{\sum\limits_{t=1}^{b} \phi(h_t)\sigma} \left[\sum_{t=1}^{b} \phi(h_t) K\left(\frac{r - r_t}{\sigma}\right) \right], \qquad r = 1, 2, \ldots, p \tag{4}$$

where K is the kernel function, and σ is the smoothing parameter called bandwidth.

The subspace dimensionality is drawn from the R distribution in this study. Again, the inversion method of the theory of pseudorandom number generation is used to implement the algorithm for determining the subspace dimensionality based on R.

1) Generate a uniform random number υ on [0, 1].
2) Determine the subspace dimensionality is r if $F_R(r-1) < \upsilon < F_R(r)$, where F_R denotes the cumulate density function of R distribution, and $1 \leq r \leq p$.

The R distribution will be updated during the construction of the B classifiers in the ensemble, and the updating process is described in Section III-C.

C. DSM

After estimating the W distribution and the R_0 distribution, the classifiers in the ensemble start being constructed. The R distribution can be automatically updated by the performance of subsequent classifier. The steps of the proposed DSM are described as follows, and the algorithm of DSM is presented in Algorithm 1.

Let B be the number of classifiers in the ensemble and the index $k = 1, 2, \ldots, B$.

1) Draw a new subspace dimensionality r'_k from R_{k-1} distribution.
2) Obtain a reduced-dimensional data set by $\widetilde{D}'_k = WBS(D, r'_k, W)$.
3) Obtain the kth component classifier of the ensemble by $h'_k = \Psi(\widetilde{D}'_k)$.
4) Estimate the resubstitution accuracy $\phi(h'_k)$ as the feedback to obtain an updating R_k distribution by

$$f_R(r) = \frac{1}{\left(\sum\limits_{t=1}^{b} \phi(h_t) + \sum\limits_{\ell=1}^{k} \phi\left(h'_\ell\right)\right)\sigma} \times \left[\sum_{t=1}^{b} \phi(h_t) K\left(\frac{r - r_t}{\sigma}\right) + \sum_{\ell=1}^{k} \phi\left(h'_\ell\right) K\left(\frac{r - r'_\ell}{\sigma}\right) \right] \tag{5}$$

5) Back to Step 1 until B classifiers have been trained.

Algorithm 1. The algorithm of DSM

```
Input:
  The training data set $D$
  The test sample $Y$
  A learning algorithm (classifier) $\Psi$
  The ensemble size $B$
  The band selection based on $W$, $WBS$
Output:
  Final hypothesis $F : Y \rightarrow c \in \{1, 2, \ldots, L\}$ computed by
  the ensemble $H = \{h'_1, h'_2, \ldots, h'_B\}$.
A. Training procedure
  Begin
    Estimate the $W$ distribution.
    Estimate the $R_0$ distribution.
    for $k = 1, 2, \ldots, B$
      Draw a subspace dimensionality $r'_k$ from $R_{k-1}$.
      $\widetilde{D}'_k = WBS(D, r'_k, W)$
      $h'_k = \Psi(\widetilde{D}'_k)$
      Obtain $R_k$ distribution by the formula (5).
    end
  End
B. Classification procedure
  $F = \arg\max_{c \in \{1,2,\ldots,L\}} card(k | h'_k(Y)) = c)$, where $k =$
  $1, 2, \ldots, B$.
```

IV. EXPERIMENTAL DESIGN

A. Methods

For investigating the multiclass classification performances of the proposed methods, there are five different algorithms used for comparison. All algorithms and their descriptions are listed in Table I. The value of b for building the initial R distribution (R_0) is set to 5, and the ensemble size B is set to 20 in RSM and DSMs. In RSM and DSMs, the simple majority voting is used for the fusion of all ensemble classifiers.

In DSMs, the kernel function K used in the R distribution is taken to be a Gaussian function as

$$K\left(\frac{t}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}. \tag{6}$$

From [17], if σ is large, then the R distribution is flatter, and the difference of selecting probabilities of bands is small. The bandwidth σ suggested by [17] is set as

$$\sigma = 0.9An^{-1/5} \tag{7}$$

where $A = \min(\text{standard deviation, interquartile range}/1.34)$, and n is the cardinality of the subspace dimensionalities that have been input.

B. Base Classifiers

To explore the performances of RSM and DSM on different base classifiers, we employ Gaussian ML classifier [25], k-nearest-neighbor classifier (kNN, $k = 1$) [25], support vector machine (SVM) [28] using a radial basis function (RBF) as a kernel, and Bayesian contextual classifier (BCC) [23] into all algorithms. The following explains why we select these classifiers as the base learners. Here, we give the term "weak classifier" a general definition that refers to a classifier that does not have a good enough performance on hyperspectral image classification with insufficient training data. We try to apply DSM with these classifiers to obtain a better performance.

The most widely used statistical classifier, namely the ML classifier, belongs to the parametric model that is made up of mean vector and covariance matrix for a normal distribution [25]. However, the covariance matrix of ML may be singular or near-singular (i.e., noninvertible) and leads to inaccurate estimation when the data dimensionality exceeds the number of training samples [29]. Consequently, the classifier performs poorly.

The kNN classifier is a simple and appealing approach, which assigns an unknown point to the class most common among its k nearest neighbors. However, high-dimensional bands obstacle to the generation of kNN since nearest neighbors of a point can be very far away, causing bias and degrading the performance of the rule [30]. Since kNN is sensitive to input bands [31], DSM generates a diverse set of kNN ensemble to overcome the mentioned problem in a reduced-dimensional space. In this study, PRTools [32] is used to implement kNN classifier.

The SVM, a successful learning algorithm commonly used for classification and regression issues, is designed by solving a constrained optimization problem. Geometrically, the SVM aims at finding a linear discriminate function with the maximal margin in the potentially very high-dimensional space. Given a training data set $D = \{(\mathbf{x}_i, c_i)\}$, where $\mathbf{x}_i \in \Re^n$, $c_i \in \{+1, -1\}$, and $i = 1, 2, \ldots, N$. The goal for SVM is to find the separating hyperplane $\mathbf{w}^T\varphi(\mathbf{x})$ that maximizes the margin, and it requires the solution of the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to} \quad c_i\left(\mathbf{w}^T\varphi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \end{aligned} \tag{8}$$

where C and ξ are penalty parameters and slack variables, respectively, for the soft-margin SVM. Using the so-called Kuhn–Tucker theorem [33] the optimization of (8) can then be reformulated as the following dual problem with respect to the Lagrange multipliers $\alpha_i \geq 0$:

$$\begin{aligned} &\min_{\alpha} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{subject to} \quad \sum_{i=1}^{N}\alpha_i c_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \\ &\qquad\qquad \forall i = 1, 2, \ldots, N \end{aligned} \tag{9}$$

where $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is called the kernel function. In this study, an RBF kernel is used as follows:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right). \tag{10}$$

The LIBSVM [34] is used to implement the SVM classifier. Here, we use the fivefold cross-validation and the grid search to find the best C within the given set $\{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ and the best γ within the given set $\{2^{-15}, 2^{-13}, \ldots, 2^{3}\}$ (suggested by Hsu *et al.* [35]) of parameters.

Although SVM has been found to provide better classification results than other widely used classifiers in hyperspectral image classification [36], [37], the band-reduction procedure combined with SVM for classification also proves its validness for obtaining higher accuracies [38], [39]. Hence, we include SVM as the base learner for investigating the effectiveness for the ensemble method.

The MRF-based BCC [23] is also applied as a base classifier. Let $u(i, j)$ denote a field that contains the classification of a pixel at the ith row and the jth column in an image X, where $u \in \{1, 2, \ldots, L\}$. According to [23], a decision rule is derived as follows:

$$\begin{aligned} u(i,j) = \mathop{\arg\max}_{u=\{1,2,\ldots,L\}} \Bigg[&-\ln\left|\sum\nolimits_u\right| + (X(i,j) - \mu_u)^T \\ &\times\Sigma_u^{-1}\left(X(i,j) - \mu_u\right) + 2m\beta + const.\Bigg] \end{aligned} \tag{11}$$

where $\sum_u$ and μ_u are the covariance matrix and mean vector of class u, respectively. The coefficient β emphasizing the significance of interaction among adjacent pixels inside a clique is empirically set to 30, and m is the total number of occurrences of the class different from $u(i, j)$ in all cliques, where MRFs are used to model the context-dependent information. The 4-neighborhood system and the corresponding cliques of order 2 are used in this study. Additionally, [23] also provides a recursive process for adaptively estimating the statistics of mean vectors and covariance matrices. In this study, we omit this step for saving the computational time.

Although the BCC can achieve satisfactory classification results [19], [23], [40], [41], it still suffers from the singular or near-singular problem in the estimation of the inverse of the covariance matrix, which makes the classifier weak and has poor classification performances. In this study, we apply BCC into the proposed DSM as the base classifier to try to overcome this problem. Additionally, we also want to investigate the

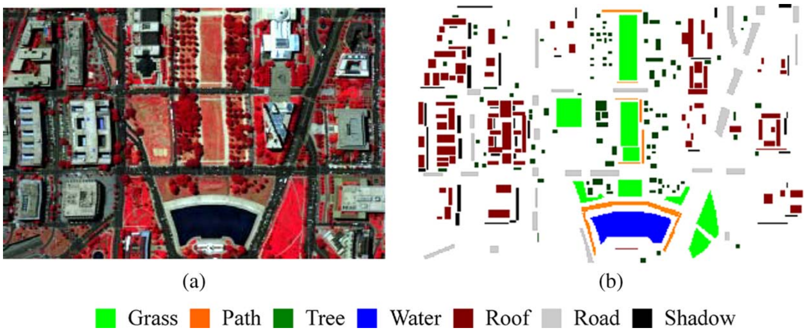


Fig. 4. (a) Test image of a portion of Washington, DC Mall data set with a size of 205 × 307 pixels. Bands 63, 52, and 36 of 191 bands were used for this image space presentation. (b) Corresponding labeled field map.

TABLE II
NUMBERS OF PIXELS IN THE WASHINGTON, DC MALL DATA SET

Class	Roof	Road	Trail	Grass	Tree	Water	Shadow	**Total**
# of pixels	3614	1982	624	2898	1446	1156	840	12560
# of training pixels	N_i	N_i	N_i	N_i	N_i	N_i	N_i	N
# of test pixels	300	300	300	300	300	300	300	2100

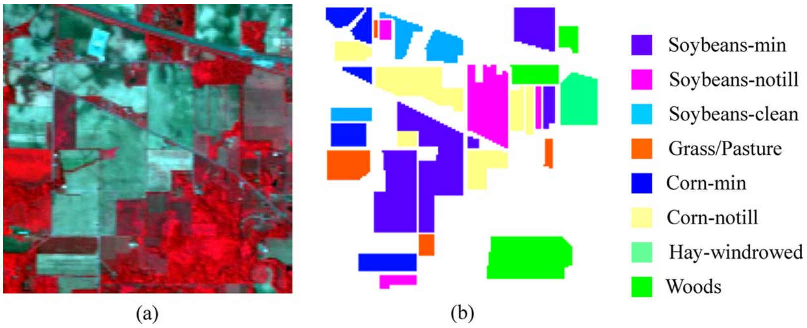


Fig. 5. (a) Test image of the Indian Pines data set. Bands 50, 27, and 17 of 220 bands were used for this image space presentation. (b) Corresponding labeled field map.

mutual effect on the behavior of selecting subspaces when using both spectral and spatial information.

C. Data Sets

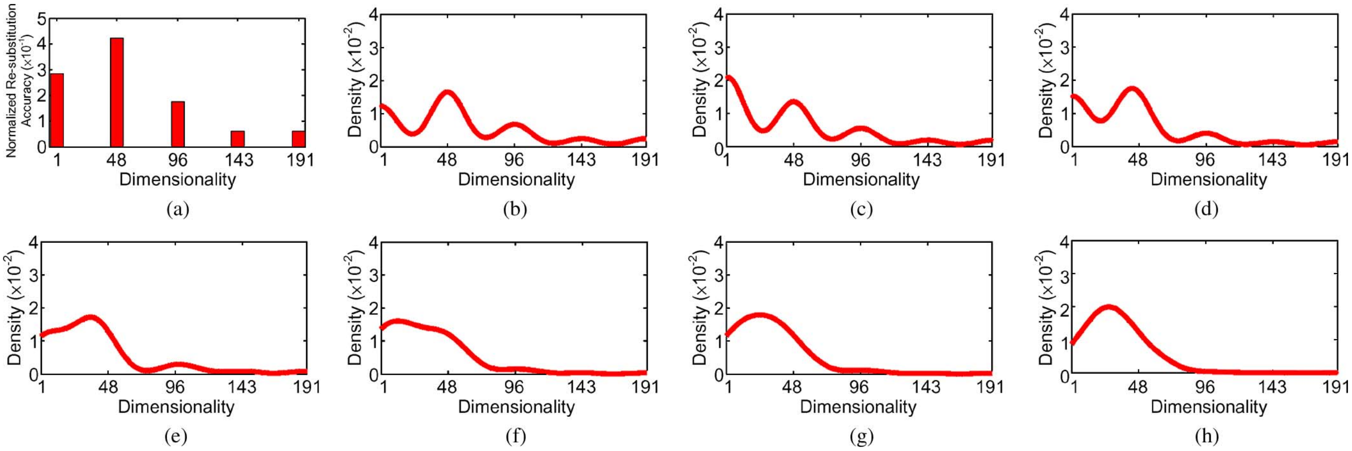
In this study, two hyperspectral image data sets are applied to compare the performances of five algorithms described in Table I. They are the urban site over Washington, DC Mall, U.S. [42] and the mixed forest/agricultural site over northwest Indiana's Indian Pines site, U.S. [43]. The first data set is a Hyperspectral Digital Imagery Collection Experiment (HYDICE) airborne hyperspectral-data flightline over Washington, DC Mall with an original size of 1280 × 307 pixels, and we use a size of 205 × 307 in our study. Two hundred and ten bands are collected in the 0.4–2.4 μm region of the visible and infrared spectrum. Some water absorption channels are discarded, resulting in 191 channels. In the experiment, seven information classes, namely, Roof, Road, Trail, Grass, Tree, Water, and Shadow, are selected by using MultiSpec [42], which is shown in Fig. 4(b), and the number of samples of each class is displayed in Table II.

For exploring the effects of the training sample size to the dimensions, three different cases, namely, $N_i = 20 < N < p$ (case 1: ill-posed problem), $N_i = 40 < p < N$ (case 2: poorly posed problem), and $p < N_i = 300 < N$ (case 3: well-posed problem), are investigated. For test sample size, we use a fixed size of 300 pixels for each class of the Washington DC Mall data set. In the Indian Pines data set, 37.24% of the labeled samples of each class is used as test samples because in training sample size $N_i = 300$ case, the maximum available test samples of the Hay-windrowed class is 178, which is 37.24% of the labeled samples. This way, smaller classes will be tested with a smaller number of pixels, and larger classes will have a larger number of samples. In each experiment, ten spatially disjointed training and test data sets are randomly assembled for estimating the parameters and computing the overall classification accuracy of the test data sets.

The Indian Pines data set is gathered by a sensor known as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). These data are obtained from an aircraft flown at 19 812 m altitude and operated by the National Aeronautics and Space Administration/Jet Propulsion Laboratory, with a size of 145 × 145 pixels and 220 spectral bands measuring approximately 20 m across on the ground. The test image is shown in Fig. 5. From the 16 different land-cover classes available in the

TABLE III
NUMBERS OF PIXELS IN THE INDIAN PINES DATA SET

Class	Corn-notill	Corn-min	Grass/ Pasture	Hay-windrowed	Soybeans-notill	Soybeans-min	Soybeans-clean	Wood	**Total**
# of pixels	1428	830	483	478	972	2455	593	1265	8504
# of training pixels	N_i	N_i	N_i	N_i	N_i	N_i	N_i	N_i	N
# of test pixels	531	309	179	178	361	914	220	471	3163



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 6. Update process of the R distribution using DSMw2 with ML classifier on the Washington, DC dataset (case 3). (a) Initialize R_0. (b) KS (R_0). (c) Iteration 1 (R_1). (d) Iteration 2 (R_2). (e) Iteration 5 (R_5). (f) Iteration 10 (R_{10}). (g) Iteration 15 (R_{15}). (h) Iteration 20 (R_{20}).

TABLE IV
AVERAGE CLASSIFICATION ACCURACY ± STANDARD DEVIATION AND KAPPA STATISTIC ± STANDARD DEVIATION OF TEN TEST DATA ON THE WASHINGTON, DC MALL DATA SET (IN PERCENT)

Algorithm	Base classifier	$N_i = 20$	$N_i = 40$	$N_i = 300$
		accuracy±standard deviation (kappa statistic±standard deviation)		
Single ML	ML	N/A	N/A	81.7±0.3 (79.9±1.0)
RSM ($r = 95$)		N/A	N/A	**93.4±0.4 (90.9±0.3)**
DSM		89.6±0.2 (88.4±1.3)	92.6±1.7 (89.5±1.4)	**94.5±0.3 (92.7±0.5)**
DSMw1		91.1±0.5 (90.8±1.2)	**94.1±0.7 (92.7±0.5)**	**94.4±0.4 (95.1±0.7)**
DSMw2		**92.1±0.6 (91.5±2.5)**	93.9±0.6 (91.5±0.4)	**95.2±0.9 (92.9±0.9)**
Single *k*NN	*k*NN	79.2±1.5 (78.6±2.0)	82.3±0.7 (82.0±0.5)	92.1±0.7 (91.6±0.7)
RSM ($r = 95$)		80.4±1.3 (77.4±1.6)	84.4±0.9 (82.7±1.7)	92.5±1.2 (90.7±1.4)
DSM		80.5±1.2 (77.8±0.6)	82.8±0.4 (81.9±0.4)	92.0±1.0 (89.6±1.0)
DSMw1		81.6±1.5 (78.3±1.2)	83.4±0.4 (82.6±0.3)	92.1±1.9 (90.9±1.3)
DSMw2		**86.8±2.2 (86.4±1.6)**	**88.4±0.7 (87.7±1.0)**	**95.9±1.2 (95.7±0.7)**
Single SVM	SVM	79.9±0.5 (76.9±1.6)	85.1±1.3 (82.1±2.0)	91.7±0.4 (88.8±0.3)
RSM ($r = 95$)		79.0±2.1 (78.2±1.6)	83.6±1.6 (82.1±0.8)	92.0±0.9 (89.1±1.6)
DSM		79.2±1.0 (76.5±0.8)	84.0±1.5 (81.5±1.2)	91.4±0.1 (88.9±0.4)
DSMw1		81.9±1.5 (78.1±1.6)	84.6±1.5 (81.9±1.6)	92.5±0.6 (89.7±0.8)
DSMw2		**87.8±0.3 (85.3±0.1)**	92.5±0.7 (90.3±1.1)	**94.5±0.2 (92.1±0.4)**
Single BCC	BCC	N/A	N/A	96.3±0.4 (94.5±0.4)
RSM ($r = 95$)		N/A	N/A	95.6±1.2 (94.1±0.9)
DSM		90.9±0.5 (88.5±0.3)	93.4±1.2 (90.7±0.5)	96.2±0.4 (94.7±0.8)
DSMw1		93.2±0.3 (90.9±0.4)	95.6±0.6 (94.8±0.7)	96.4±0.6 (95.7±1.0)
DSMw2		**94.4±1.4 (93.4±1.3)**	**95.4±0.7 (93.9±1.3)**	**97.0±0.6 (95.8±0.7)**

TABLE V
AVERAGE CLASSIFICATION ACCURACY ± STANDARD DEVIATION AND KAPPA STATISTIC ± STANDARD DEVIATION OF TEN TEST DATA ON THE INDIAN PINES DATA SET (IN PERCENT)

Algorithm	Base classifier	$N_i = 20$	$N_i = 40$	$N_i = 300$
		accuracy±standard deviation (kappa statistic±standard deviation)		
Single ML	ML	N/A	N/A	71.2±0.7 (66.0±0.7)
RSM ($r = 110$)		N/A	N/A	84.2±4.5 (81.2±5.3)
DSM		67.6±3.4 (62.7±3.7)	72.7±0.9 (67.9±1.1)	86.9±1.3 (84.4±1.6)
DSMw1		66.7±2.8 (60.8±3.0)	74.7±2.0 (70.0±2.3)	85.8±3.4 (83.2±4.0)
DSMw2		**68.7±3.0 (63.2±3.3)**	**75.4±1.7 (71.0±1.9)**	**87.2±0.8 (84.8±0.9)**
Single *k*NN	*k*NN	59.7±2.6 (52.7±2.9)	**66.5±1.1 (60.7±1.1)**	83.0±0.6 (79.8±0.7)
RSM ($r = 110$)		60.6±1.5 (53.6±1.6)	65.6±1.7 (59.7±1.9)	84.4±0.8 (81.9±0.9)
DSM		60.4±1.5 (53.6±1.6)	66.2±1.1 (60.4±1.1)	83.9±0.7 (80.9+±0.9)
DSMw1		60.3±1.5 (53.7±1.4)	66.2±1.5 (60.4±1.7)	83.8±0.6 (80.8±0.7)
DSMw2		**60.8±0.9 (54.0±0.9)**	66.4±1.2 (60.6±1.3)	**85.0±0.5 (82.2±0.6)**
Single SVM	SVM	74.1±2.1 (69.4±2.5)	77.3±1.4 (73.3±1.6)	82.8±0.7 (79.6±0.8)
RSM ($r = 110$)		76.6±1.6 (72.4±1.9)	81.4±1.6 (78.0±1.9)	89.6±0.3 (87.6±0.4)
DSM		76.4±2.1 (72.2±2.4)	81.5±1.7 (78.2±2.0)	**90.2±0.7 (88.3±0.9)**
DSMw1		76.9±1.7 (72.7±2.0)	**81.9±1.6 (78.6±1.9)**	89.9±1.1 (88.0±1.2)
DSMw2		**77.1±1.8 (73.0±2.0)**	81.8±1.5 (78.4±1.8)	89.5±0.8 (87.5±0.9)
Single BCC	BCC	N/A	N/A	83.3±1.0 (80.2±1.1)
RSM ($r = 110$)		N/A	N/A	93.1±2.4 (91.8±2.8)
DSM		72.2±3.9 (67.2±4.6)	**80.4±1.6 (76.8±1.8)**	94.6±1.6 (93.5±1.9)
DSMw1		71.4±3.0 (66.4±3.3)	80.1±2.4 (76.6±2.7)	93.3±3.3 (92.0±2.0)
DSMw2		**73.1±4.1 (68.2±4.6)**	78.0±1.4 (74.1±1.6)	**96.2±0.5 (95.5±0.6)**

original ground truth [42], eight are discarded due to the constraint of three sample sizes. The eight classes, namely, Corn-notill, Corn-min, Grass/Pasture, Hay-windrowed, Soybeans-notill, Soybeans-min, Soybean-clean, and Wood, are selected for the experiments, and the number of samples of each class is displayed in Table III.

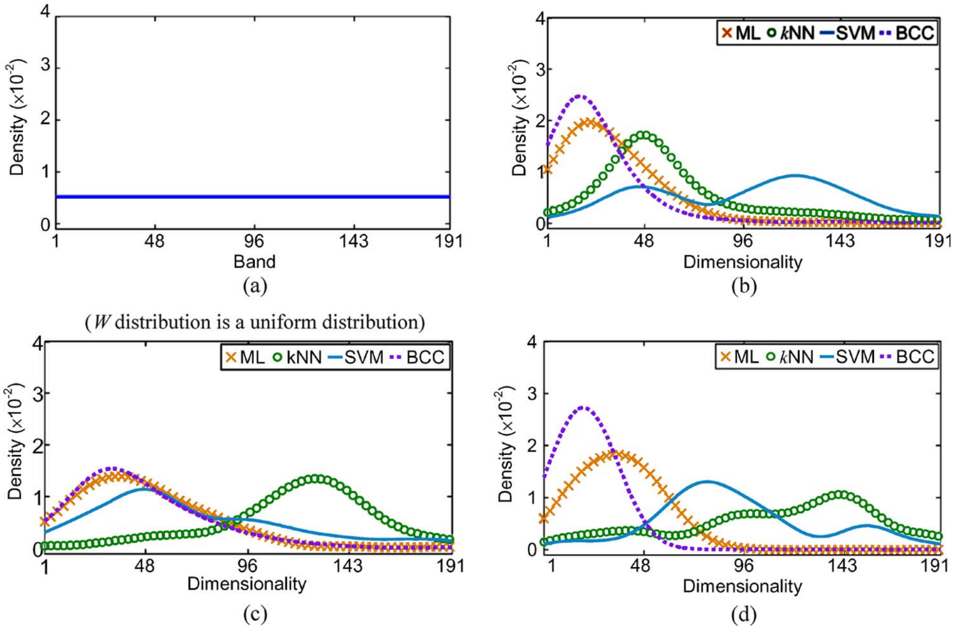
V. EXPERIMENT RESULTS

Fig. 5 demonstrates the update process of R distribution using DSMw2 with ML classifier on the Washington DC data set. Initially, R starts from five specific subspace sizes, namely, 1, 48, 96, 143, and 191, with approximately equal intervals as Fig. 6(a), and then, kernel density estimation is introduced to form R_0 as the first guide to select the first subspace size as Fig. 6(c)–(h) shows the change of R distribution and the corresponding subspace size selection are based on these distributions. Through several times of updates, the R distribution for subspace size selection tends to be stable.

Tables IV and V display the classification accuracies of testing data with cases 1, 2, and 3 on the Washington, DC Mall and Indian Pines data sets, respectively. Note that the shaded parts indicate the best accuracy of each case, and the best accuracy of each applied classifier among all algorithms is written in bold type in accordance to each case. Figs. 7–12 are three types of W distributions and corresponding R distributions of two data sets. Note that all R distributions are the final result over 20 iterations. The following are some findings based on these results.

A. *Washington, DC Mall Data Set*

1) In the Washington, DC data set, the highest accuracies among all methods are 94.4%, 95.4%, and 97.0% in cases 1, 2, and 3, respectively, and all occur in DSMw2 with BCC. Additionally, as the training sample size increases, the accuracies also represent ascending tendencies in all combinations.
2) In terms of each classifier, the best accuracies occur mostly when applying DSMw2 among three cases. Additionally, the proposed methods, namely, DSM, DSMw1, and DSMw2, are better than single classifiers and RSM disregarding any base classifier applied.
3) The W distribution in Fig. 9 is significantly different from those in Figs. 7 and 8, possibly giving sound results of DSMw2. This shows that using the LDA separability as the band weights is a better choice to select component bands of subspaces.
4) The W distribution of kNN in Fig. 8(a) shows a different behavior with respect to the three others and is closer to a uniform distribution. It is possible to be the result of the similar classification accuracies between DSM and DSMw1.



(*W* distribution is a uniform distribution)

Fig. 7. (a) W distribution and (b)–(d) corresponding R distributions of DSM using ML, kNN, SVM, and BCC, respectively, on the Washington, DC Mall dataset: (a) RBS (W distribution is a uniform distribution); (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).

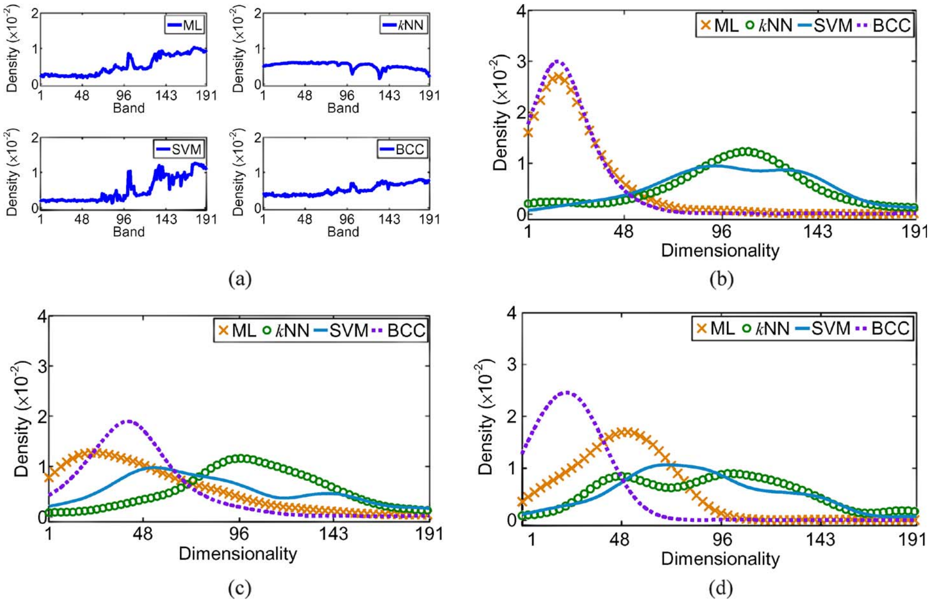


Fig. 8. (a) W_{ACC} distribution and (b)–(d) corresponding R distributions of DSMw1 using ML, kNN, SVM, and BCC, respectively, on the Washington, DC Mall dataset: (a) W_{ACC} distributions; (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).

5) RSM is mostly better than single classifiers. ML and BCC do not work in cases 1 and 2 under RSM due to the singularity problem. DSM can reduce the singularity problem.
6) From R distributions of Figs. 7–9, the proposed method will automatically estimate R distributions for different base classifiers. Based on the dimensionality of subspace, lower dimensionality is suitable for ML and BCC, whereas higher dimensionality is suitable for kNN and SVM. Furthermore, due to applying additional spatial information, BCC definitively uses less dimensionality than ML.

B. Indian Pines Data Set

1) The highest accuracies among all methods are 77.1% (DSMw2 with SVM), 81.9% (DSMw1 with BCC), and 96.2% (DSMw2 with BCC) in cases 1, 2, and 3, respectively.
2) The best accuracies are distributed over DSM, DSMw1, and DSMw2 when using SVM and BCC. For kNN and SVM classifiers, the performances of three DSMs seem to be similar to that of RSM, which means that half of the original space size is suitable for RSM with kNN and SVM classifiers. The R distributions in Figs. 10–12 support this claim; more importantly, they demonstrate

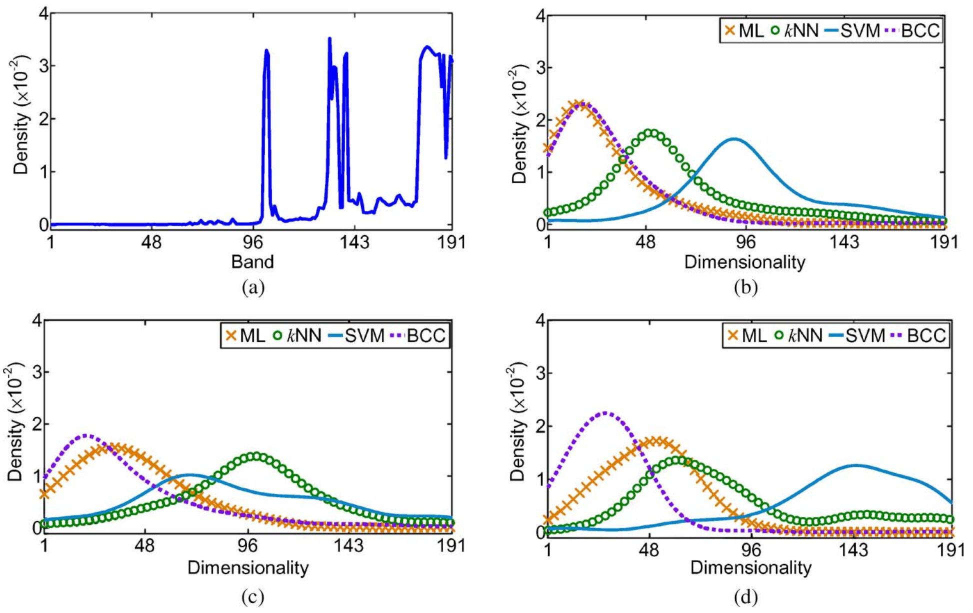


Fig. 9. (a) W_{LDA} distribution and (b)–(d) corresponding R distributions of DSMw2 using ML, kNN, SVM, and BCC, respectively, on the Washington, DC Mall dataset: (a) W_{LDA} distribution; (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).

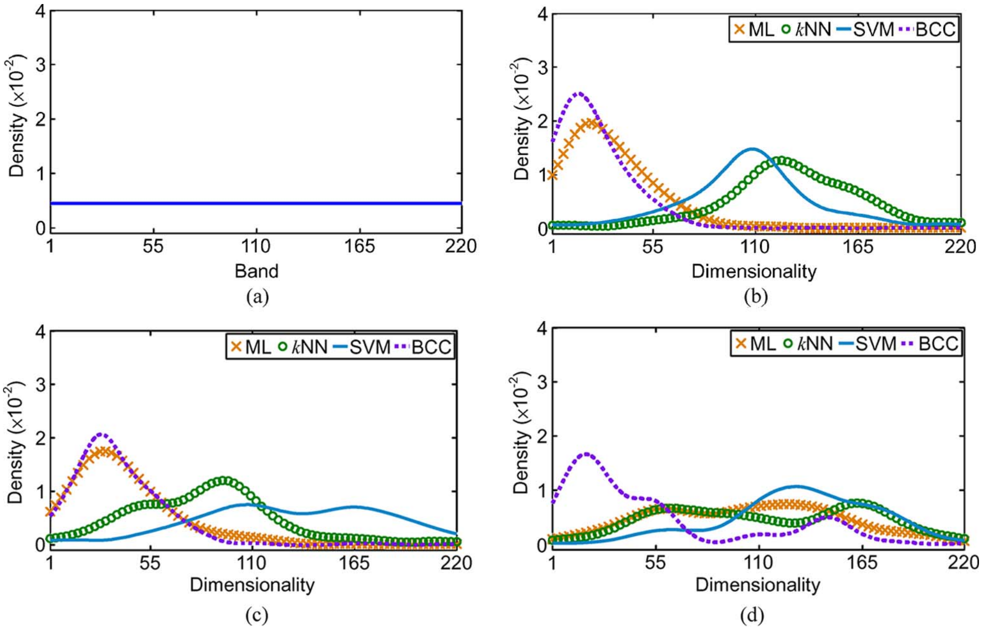


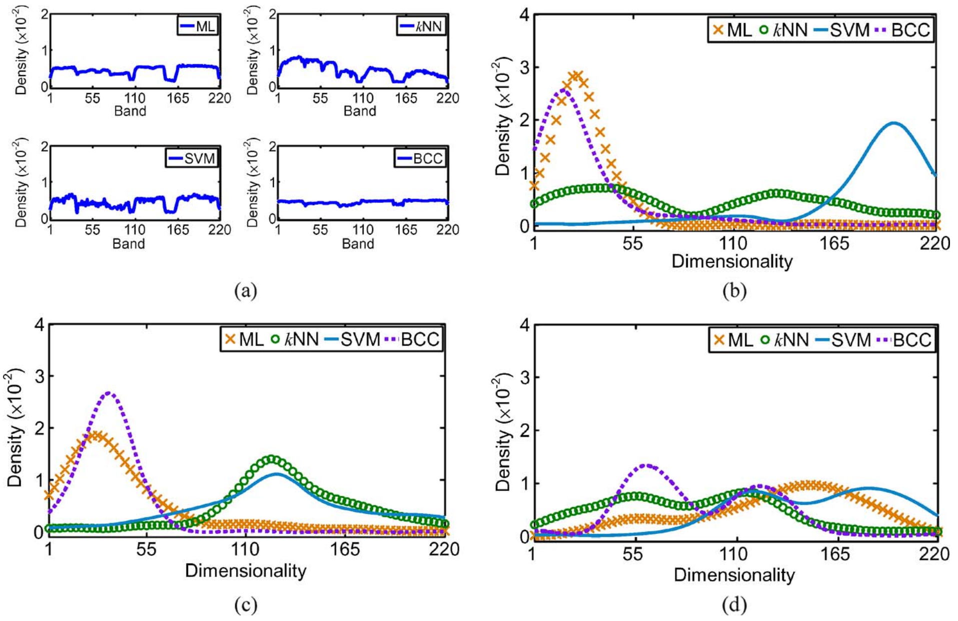
Fig. 10. (a) W distribution and (b)–(d) corresponding R distributions of DSM using ML, kNN, SVM, and BCC, respectively, on the Indian Pines dataset: (a) RBS (W distribution is a uniform distribution); (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).

that the suitable subspace size for kNN and SVM classifiers is close to half of the original space size, which matches Ho's suggestion. Additionally, these R distributions reveal that BCC uses less dimensionality than other classifiers.

3) In cases 1 and 2, ML and BCC suffer from the singular problem when the dimensionality of subspace exceeds the training sample size. The proposed dynamic selection scheme can avoid this situation.

4) In Figs. 7–9, the W distributions are dissimilar; therefore, the performances of DSM, DSMw1, and DSMw2 are different as well. In Figs. 10–12, the W distributions are flat and similar; therefore, the performances of DSM, DSMw1, and DSMw2 are close as well.

Due to length constraints, only some classified images are shown for comparison and three methods (single classifier, RSM, and DSMw2) are selected to generate the classified images under case 3. Figs. 13–15 are the classification results of the area of Fig. 4 using single classifier, RSM and DSMw2 with ML, kNN, SVM, and BCC, respectively. Generally, we can find that all single classifiers do not perform well compared to RSM and DSMw2. In Fig. 13, although BCC shows less speckle error than other classifiers, there are many pixels from roads that are incorrectly identified as roofs. Compared to the



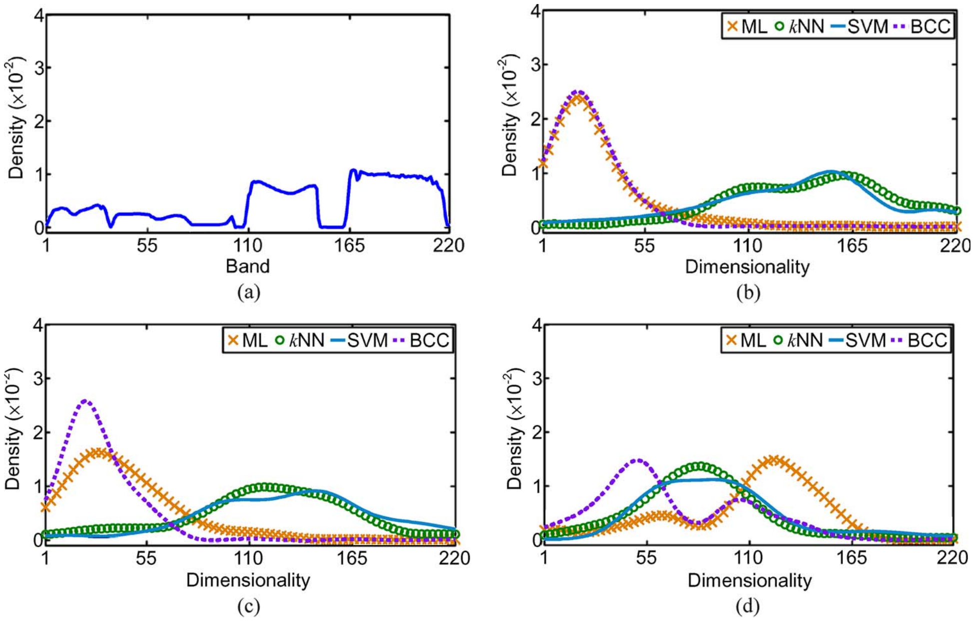
(a)

(b)

(c)

(d)

Fig. 11. (a) W_{ACC} distribution and (b)–(d) corresponding R distributions of DSMw1 using ML, kNN, SVM, and BCC, respectively, on the Indian Pines dataset: (a) W_{ACC} distributions; (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).



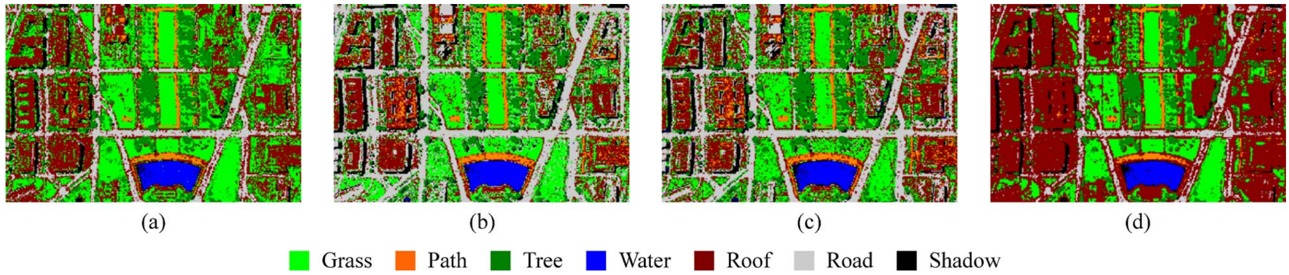
(a)

(b)

(c)

(d)

Fig. 12. (a) W_{LDA} distribution and (b)–(d) corresponding R distributions of DSMw2 using ML, kNN, SVM, and BCC, respectively, on the Indian Pines dataset: (a) W_{LDA} distributions; (b) case 1 ($N_i = 20 < N < p$); (c) case 2 ($N_i = 40 < N < p$); and (d) case 3 ($p < N_i = 300 < N$).



(a)

(b)

(c)

(d)

Fig. 13. Thematic maps resulting from the classification of Fig. 4 in case 3: (a)–(d) are the results of the single classifier. (a) ML. (b) kNN. (c) SVM. (d) BCC.

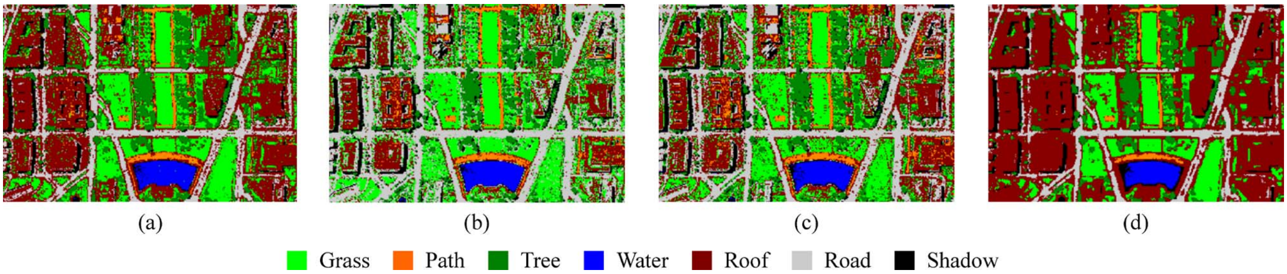


Fig. 14. Thematic maps resulting from the classification of Fig. 4 in case 3: (a)–(d) are the results of using RSM. (a) ML. (b) kNN. (c) SVM. (d) BCC.

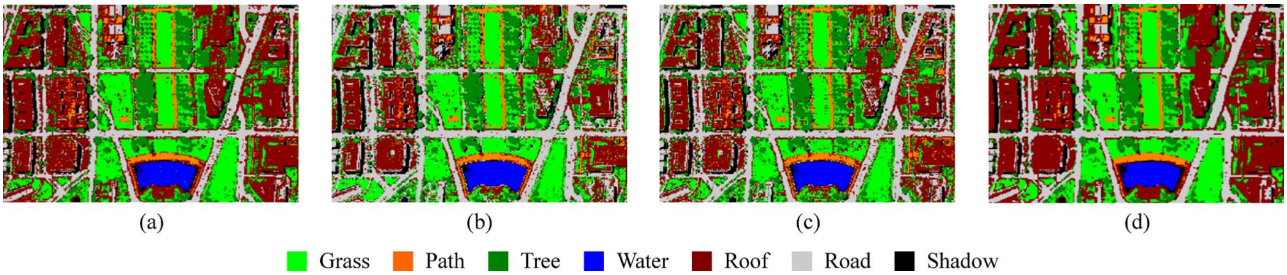


Fig. 15. Thematic maps resulting from the classification of Fig. 4 in case 3: (a)–(d) are the results of using DSMw2. (a) ML. (b) kNN. (c) SVM. (d) BCC.

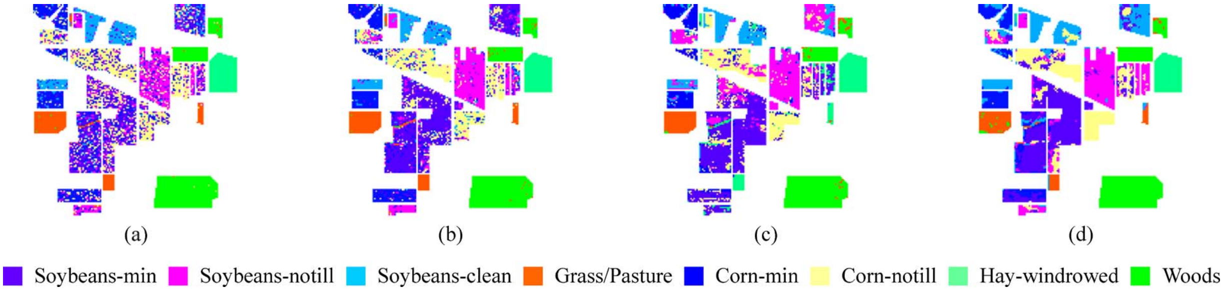


Fig. 16. Thematic maps resulting from the classification of Fig. 5 in case 3: (a)–(d) are the results of the single classifier. (a) ML. (b) kNN. (c) SVM. (d) BCC.

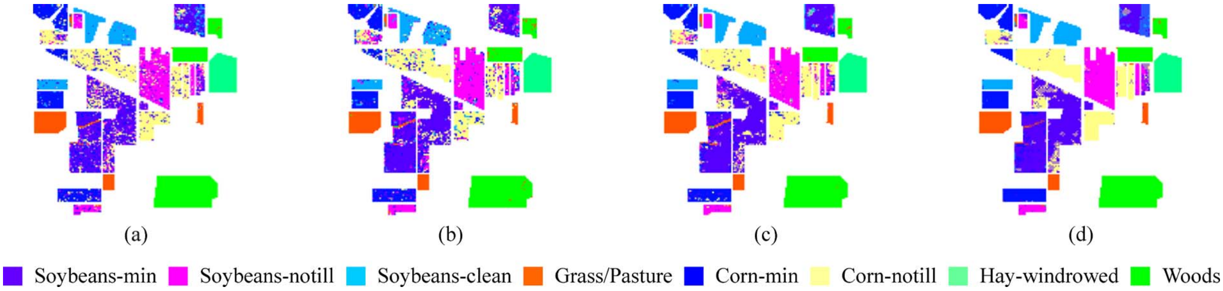


Fig. 17. Thematic maps resulting from the classification of Fig. 5 in case 3: (a)–(d) are the results of using RSM. (a) ML. (b) kNN. (c) SVM. (d) BCC.

single classifier method, RSM and DSMw2 both obtain improvement in roofs; DSMw2 with kNN and SVM significantly outperform the single classifier method and RSM in grass. The best classification result occurs in Fig. 15(d) by using DSMw2 with BCC.

Figs. 16–18 are the classification results of the area of Fig. 5 using single classifier, RSM and DSMw2 with ML, kNN, SVM, and BCC, respectively. Compared to the ground truth in Fig. 5(b), we can observe that the classification results of RSM and DSMw2 are better than those of the single classifiers, particularly in Soybeans-min, Soybeans-notill, and Corn-notill, which are the most difficult parts to accurately classify; additionally, RSM and DSMw2 have similar performances when using ML, kNN, and SVM. The best classification result occurs in Fig. 18(d) by using DSMw2 with BCC, which performs much better than RSM with BCC in Soybeans-min.

VI. Conclusion and Comments

In this paper, a new multiple classifiers system named DSM has been proposed for classifying hyperspectral image data, and we have investigated the effects of using four different base classifiers and three training sample sizes. Compared to

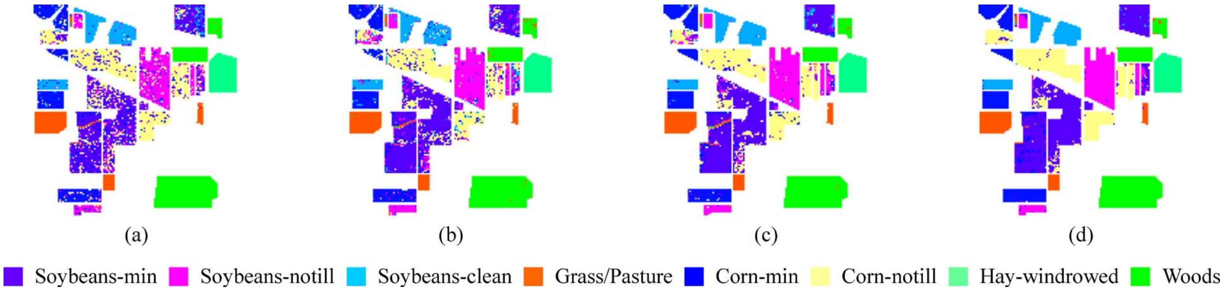


Fig. 18. Thematic maps resulting from the classification of Fig. 5 in case 3. (a)–(d) Results of using DSMw2. (a) ML. (b) kNN. (c) SVM. (d) BCC.

original RSM, DSM shows its statistical foundation for selecting better subspaces and their sizes, at the same time having the robust ability to accommodate every situation within this study.

In the original RSM, the probability of each band being selected is based on uniform distribution, whereas it is replaced by W distribution in DSM. Two criteria, namely, the resubstitution accuracy and the separability of Fisher's LDA, are used to model the density of W distribution. Two "dynamic" strategies are carried in the R distribution. One is that the R distribution is enabled to automatically select suitable subspace sizes, which are usually troublesome to preprocess. The other one is the updating technique, which makes the R distribution change progressively toward a stable status. Experimental results show that these modifications have the ability to improve the classification accuracy.

There are two theoretical drawbacks to the proposed DSM. The first one is that in the estimation of the W distribution, the accuracy of the resubstitution classifications is actually evaluated using only a single band. Theoretically, this approach does not precisely measure the importance of each band mentioned previously because there is cross-information between bands. In terms of algorithm, once the dimensionality of subspace has been estimated, the importance of each band should be evaluated statistically by trying different sets with the dimensionality estimated and including the band to be estimated. However, there are many combinations of band sets; therefore, the computation load will increase to obtain a better W.

The second drawback is that in the estimation of R distribution, the classification performances of b classifiers built in b different-dimensional spaces $(r_1, \ldots, r_b)$ are adopted, which means that for each r_i dimensional space, only one classifier is trained. The results could be unstable because the data set extracted may be unrepresentative. The methods to alleviate this problem are to create many sets with identical dimensionality and perform the single classifier on each set, then averaging the results, or to use a sort of k-fold validation.

As the two approaches are applied to W and R, they may yield better results but may greatly increase the computational load. However, we have experimentally found that the proposed DSM, indeed, yields better results when B is smaller than 15, but the results of the two approaches tend to be similar when B is bigger than 15. By using the proposed method, computation time is reduced and yields similar results. Due to length constraints, experimental results are not presented.

In conclusion, the proposed DSM fixes the inadequacies of RSM by employing two importance distributions in the process of subspace selection, and furthermore, it not only alleviates the Hughes effect but also obtains sound results in classification performance. The experimental results also show that DSMw2 with BCC has the best performance in accuracy and classification map. An interesting finding from the R distribution shows that BCC performs well in a much smaller dimensional space. Comparing the performances of DSM with ML and BCC, we find that the spatial information does, in fact, improve DSM.

Acknowledgment

The authors would like to thank Dr. Landgrebe for providing the Indian Pines site and Washington, DC mall data sets and the reviewers for their constructive comments.

References

[1] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, Jun. 2002.

[2] L. Breiman (1998, Jun.). Arcing classifiers. *Ann. Stat.* [Online]. *26(3)*, pp. 801–824. Available: http://www.jstor.org/stable/120055

[3] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[4] R. E. Bellman, *Adaptive Control Processes—A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.

[5] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[6] T. K. Ho, "Nearest neighbors in random subspaces," in *Proc. IAPR, Lecture Notes in Computer Science: Advances in Pattern Recognition*, 1998, pp. 640–648.

[7] M. Skurichina and R. P. W. Duin, "Bagging and the random subspace method for redundant feature spaces," in *Proc. 2nd Int. Workshop Multiple Classifier Syst.*, 2001, pp. 1–10.

[8] B.-C. Kuo, C.-H. Pai, T.-W. Sheu, and G.-S. Chen, "Hyperspectral data classification using classifier overproduction and fusion strategies," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, vol. 5, pp. 2937–2940.

[9] A. Bertoni, R. Folgieri, and G. Valentini, "Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancies," in *Proc. 15th Italian Workshop Neural Nets*, Perugia, Italy, 2004, pp. 29–35.

[10] A. Bertoni, R. Folgieri, and G. Valentini, "Random subspace ensembles of support vector machines," in *Neurocomputing*, vol. 63, pp. 535–539, 2005.

[11] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[12] S. Sun, C. Zhang, and D. Zhang, "An experimental evaluation of ensemble methods for EEG signal classification," *Pattern Recognit. Lett.*, vol. 28, no. 15, pp. 2157–2163, Nov. 2007.

[13] J. W. Christopher, "Hyperspectral image classification with limited training data samples using feature subspaces," *Proc. SPIE*, vol. 5425, pp. 170–181, 2004.

[14] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[15] C.-H. Chuang, B.-C. Kuo, and H.-P. Wang, "Fuzzy fusion method for combining small number of classifiers in hyperspectral image classification," in *Proc. 8th Int. Conf. Intell. Syst. Des. Appl.*, 2008, vol. 1, pp. 26–28.

[16] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[17] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1985.

[18] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.

[19] R. Cossu, S. Chaudhuri, and L. Bruzzone, "A context-sensitive Bayesian technique for the partially supervised classification of multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 3, pp. 352–356, Jul. 2005.

[20] L. O. Jimenez, J. L. Rivera-Medina, E. Rodriguez-Diaz, E. Arzuaga-Cruz, and M. Ramirez-Velez, "Integration of spatial and spectral information by means of unsupervised extraction and classification for homogenous objects applied to multispectral and hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 844–851, Apr. 2005.

[21] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marıacute;, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[22] X. Jia and J. A. Richards, "Managing the spectral–spatial mix in context classification using Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 311–314, Apr. 2008.

[23] Q. Jackson and D. A. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.

[24] Y. Maghsoudi, A. Alimohammadi, M. J. Valadan Zoej, and B. Mojaradi, "Application of feature selection and classifier ensembles for the classification of hyperspectral data," in *Proc. 26th Asian Conf. Remote Sens.*, Hanoi, Vietnam, 2005.

[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA: Academic, 1990.

[26] L. Devroye, *Non-Uniform Random Variate Generation*. New York: Springer-Verlag, 1986.

[27] F. van der Heiden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Chichester, U.K.: Wiley, 2004.

[28] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[29] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 763–767, Jul. 1996.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[31] S. D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intell. Data Anal.*, vol. 3, no. 3, pp. 191–209, Sep. 1999.

[32] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. M. J. Tax, and S. Verzakov, *PRTools4.1, A Matlab Toolbox for Pattern Recognition*. Delft, The Netherlands: Delft Univ. Technol., 2007.

[33] F. J. Kampas, "Tricks of the trade: using reduce to solve the Kuhn–Tucker equations," *Mathematica J.*, vol. 9, pp. 686–689, 2005.

[34] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[35] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*, 2004. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[36] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[37] L. Bruzzone and C. Persello, "A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2142–2154, Jul. 2009.

[38] B.-C. Kuo and K.-Y. Chang, "Feature extractions for small sample size classification problem," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 756–764, Mar. 2007.

[39] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2008.

[40] Y. Jhung and P. H. Swain, "Bayesian contextual classification based on modified M-estimates and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 67–75, Jan. 1996.

[41] A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.

[42] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[43] *AVIRIS NW Indiana's Indian Pines 1992 Data Set*, [Online]. Available: ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C, (original files) and ftp://ftp.ecn.purdue.edu/biehl/PC_MultiSpec/ThyFiles.zip, (ground truth)

Jinn-Min Yang received the B.S. and M.S. degrees from the National Taichung Teachers College, Taichung, Taiwan, in 1994 and 2000, respectively. He is currently working toward the Ph.D. degree in the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan.

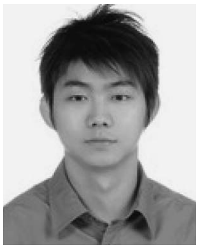
He is also with the Department of Mathematics Education, National Taichung University, Taichung. His research interests include pattern recognition, remote sensing, and machine learning.

Bor-Chen Kuo received the B.S. and M.S. degrees from National Taichung Teachers College, Taichung, Taiwan, in 1993 and 1996, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, in 2001.

He is currently a Professor with the Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taichung. His research interests include pattern recognition, remote sensing, image processing, and nonparametric functional estimation.

Pao-Ta Yu (S'88–M'90) received the B.S. degree in mathematics from the National Taiwan Normal University, Taipei, Taiwan, in 1979, the M.S. degree in computer science from the National Taiwan University, Taipei, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1989.

Since 1990, he has been with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, where he is currently a Professor. His research interests include e-learning, neural networks and fuzzy systems, nonlinear filter design, intelligent networks, and XML technology.



Chun-Hsiang Chuang received the B.S. degree from Taipei Municipal Teachers College, Taipei, Taiwan, in 2004, and the M.S. degree from the National Taichung University, Taichung, Taiwan, in 2009. He is currently working toward the Ph.D. degree in the Institute of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan.

He is also with the Brain Research Center, National Chiao-Tung University. His research interests include pattern recognition, remote sensing, and biomedical signal processing.